LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-BOOK-423388

Inferential validation and evidence interpretation

Stephan P. Velsko

February 9, 2010

Microbial Forensics

Disclaimer

Inferential validation and evidence interpretation

Stephan P. Velsko*
Lawrence Livermore National Laboratory
December 15, 2009

*Corresponding author
velsko2@llnl.gov
Ph: 925-423-0191
Fax: 925-422-8020

Running head title: Inferential validation ….

Summary

Inferential validation is a research activity that aims to provide statistical measures for the support that primary forensic measurements or observations provide for an interpretation of that data. It is a distinct research and development activity that is separable from other kinds of validation. The purposes of most microbial forensic assays can be formulated as hypothesis tests, and the support that measurements provide for a hypothesis can be expressed as a likelihood ratio. The likelihood ratio can be estimated from a receiver-operating characteristic (ROC) curve that is obtained from performing the assay on a representative sample of the "population" of samples relevant to the assay. This chapter provides some guidelines for defining the population and performing unbiased sampling, focusing on chemical and physical analysis methods.

Key Words: Interpretation, Validation, Inference, Daubert, ROC, Likelihood, Hypothesis

1. The need to validate the interpretation of microbial forensic evidence

The field of microbial forensics is being created at a time when forensic science *in general* faces unprecedented skepticism. The foundations of many long-accepted forensic science methods have been questioned, and recent National Research Council studies have supported these criticisms [1,2]. It is likely that in future cases both the admissibility and evidentiary weight of microbial forensic evidence will be closely scrutinized, and that Daubert challenges will occur. Thus, it is imperative that this new area of forensic science build sound, "Daubert resistant" foundations by carefully considering both the framework for validation and the way in which microbial forensic evidence is conveyed in reports, hearings and trials.

This concern is generally appreciated by the community of scientists engaged in microbial forensic research and operations, who have addressed certain important aspects of validation. In particular, guidelines for Quality Assurance have been formulated and published widely [3]. It is also possible to find clear and useful guidance for establishing the precision and accuracy of a variety of assays of use in microbial forensics. However, as we will show below, this addresses only one aspect of validation, and by itself cannot impart "Daubert resistance" to microbial forensic evidence. This is because the most salient criticisms that have been leveled at forensic science do not question data quality, but rather *interpretation*. This issue is best illustrated by two quotations that clearly differentiate between the validity of data, and the validity of the interpretation of that data in forensic science testimony:

> "Even if an instrument yields exquisitely precise measurements, the witness's inferences from the measurements may be badly flawed. As Justice Blackmun stressed in *Daubert*, it is the expert's ultimate inference which 'must be derived by the scientific method … [and] supported by appropriate validation…'"
>
> - Edward J. Imwinkelried in *The Methods of Attacking Scientific Evidence [4]*

> "The committee found the analytical technique used is suitable and

> reliable for use in court, as long as FBI examiners apply it uniformly as recommended. […] However, for legal proceedings, the probative value of these findings and how the probative value is conveyed to a jury remains a critical issue."
>
> - From the NRC report *Forensic Analysis: Weighing Bullet Lead Evidence [1]*

This chapter suggests that the validation of interpretation is a distinct research and development activity that is separable from other kinds of validation, and can be formalized to a large extent. Although there is a large body of literature that discusses methods for validating data interpretation, the concept is seldom treated as a separate activity in the development of forensic assays. The statistical concepts and methodologies described in this chapter play much more familiar roles in the area of medical diagnostics, where they may be considered "mainstream."

Because the term validation is used in various ways in the forensic context, a short description of the various types of validation that have been described, their inter-relationships, and their connections to the Daubert decision [5] and Federal Rules of Evidence [6] is provided in section 2. Next, a general scheme for inferential validation in the context of microbial forensics is described, followed by a discussion of population and sampling issues that apply to the specific area of chemical and physical analysis of biological agents.

2. The taxonomy of validation

The term validation is used to describe a number of distinct activities in forensic research and operations. To understand the relationships among the different classes of validation activities it is useful to turn to the original text of the *Daubert* decision, which noted that scientific validity rests on two factors: reliability and relevance [5]. The reliability of a technique is its ability to produce consistent, objective results with known precision and traceable accuracy. Those quality assurance procedures that assure the reliability of scientific evidence are termed *analytical validation*.

The term relevance refers to the fact that analytical measurements or other scientific data usually are not of interest to the court *per se*, but are proffered as evidence to support or refute *by inference* a fact at issue in the trial [4]. According to the Federal Rules of Evidence 401 and 402, relevant evidence is that "having any tendency to make the existence of any fact [at issue in the trial] more probable or less probable than it would be without the evidence." Evidence that is not relevant in this sense is not admissible. *Inferential validation* is the process that establishes the strength of support (i.e. the degree of relevance, or probative value) that a given observation or other data provide for the *expert's ultimate inference* for which the observation or data are offered as evidence. (The term ultimate inference should not be confused with the legal term *ultimate issue*, which refers to a question that the jury must decide, e.g. the guilt or innocence of the accused.) Table 1 provides some examples of assays and the ultimate inferences that they may be used to support in the field of microbial forensics.

Corresponding to these two classes of validation, it is sometimes useful to distinguish between a "reporting expert witness" and an "interpreting expert witness"[4]. The former testifies as to the test result and how it was obtained, and seeks to assure the factfinders that the results of the analysis are reliable. The latter provides an expert opinion regarding the "ultimate inference" to be drawn from that test result. (In practice, of course, the same expert may perform both these roles, and the distinction is useful even if the evidence is never used in court.) The reporting expert comes armed with the results of analytical validation, while the interpreting expert supports his testimony with the experimental results that provide inferential validation. No matter how exacting the QA and QC regime, the expert's *interpretation* of scientific evidence is still vulnerable to challenge - especially if the interpretation is particularly crucial to the prosecution narrative. For either expert, the ultimate product of a validation study is, in essence, the value of a statistical estimator. For analytical validation examples of such estimators are precision values associated with measurements, or a distribution of difference values relative to a known standard [7,8]. For inferential validation, typical estimators are ROC curves or likelihood ratio estimators for a well-constructed hypothesis test [9]. The above considerations are summarized in Table 2.

Inferential validation is intrinsically a research activity, but there are variants of analytical validation that apply to the operational implementations of forensic assays. For example, a distinction can be made between *developmental* and *internal* validation. Following the definitions given in the SWGMGF Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work [3], developmental validation is an activity carried out by the laboratory that develops the technique and thus is a research activity, while internal validation is carried out by laboratories that are implementing the technique in-house for operational use. Inferential validation need only be performed in developmental mode, since the same inferential power will apply to the same technique applied in a different laboratory, assuming that internal analytical validation has been performed.

Two other terms sometimes found in the literature are *external validation* and *preliminary validation*. In external validation, the performance of a technique by a laboratory is evaluated by one or more (usually blind) tests administered by an independent entity. In this regard, external validation is a species of analytical validation that provides additional assurance of the consistency and reliability of a technique by showing it to be independent of the particular laboratory or operator. In Microbial Forensics, preliminary validation has been defined as the acquisition of limited test data to enable the evaluation of a method used to assess materials derived from a biocrime or bioterrorism event [3,10,11]. Preliminary validation enables the evaluation of a previously uncharacterized method used to provide investigative support (e.g. generating investigative leads.) Preliminary validation involves <u>both</u> analytical and inferential validation. The latter is clearly required at some level in order to evaluate the value of the test for generating investigative leads based on the ultimate inference drawn from the test. The SWGMGF Guidelines for Microbial Forensics stipulate that if the results are to be used for other than investigative support, then a panel of peer experts, external to the laboratory, should be convened to assess the utility of the method and to define the limits of interpretation and conclusions drawn [3,11]. Table 3 summarizes the matrix of requirements for validation corresponding to the foregoing discussion.

3. The ROC/LR framework for inferential validation

Human DNA analysis is generally regarded as the "gold standard" for forensic science, and the statistical foundation for DNA evidence is sometimes suggested as a paradigm for inferential validation of other kinds of forensic tests and assays [12]. For example, the NRC studies of forensic science have consistently advocated a *likelihood* or *likelihood ratio* framework for interpreting scientific evidence [1,13,14]. (Counter-arguments to this notion are sometimes put forth by professional forensic scientists, who argue that other kinds of forensic assays and tests cannot be treated in the same framework [15].) In this section I will outline a foundation for inferential validation based on a likelihood ratio approach.

The standard likelihood equation is shown in Fig 1. E represents some piece of evidence, in our case some measurement or set of observations made on one or more samples of a biological agent. H is some hypothesis concerning the production or source of that agent. $O_0(H)$ are the odds that H is true in the absence of E, and O(H|E) are the posterior odds. The likelihood ratio is determined by the probability that E would be observed if H were true *versus* if it were false ($\bar{H}$). The likelihood ratio is often considered the strength (or probative value) of the evidence E with respect to the hypothesis H. Since Federal Rule 401 explicitly defines the relevance of evidence in terms of whether it makes H more probable or less probable, legal scholars have often cited the likelihood ratio (LR) as a measure of relevance, and hence admissibility [16-18]. Specifically, if

$$LR(E) = P(E|H)/P(E|\bar{H}) = 1,$$

the evidence is not logically relevant and thus inadmissible according to Rule 402.

Given this correspondence, the approach to demonstrating the probative value of a given test or assay with respect to a given hypothesis (e.g. those in table 1) is to estimate the likelihood ratio associated with the measurements or observations E produced by the test

when it is applied to samples that conform to H and $\bar{H}$. When the test is applied to a questioned sample and the result E is obtained, if LR(E) >1, E supports the hypothesis H; if LR(E) <1, E supports $\bar{H}$. Thus, a scientist may testify that his/her measurement of a certain value of some metric for a sample provides a particular level of support to the hypothesis in question, rather than stating that his values are "consistent with" the hypothesis (which is simply the statement that $P(E|H) \neq 0$), or worse, that the results make it "likely that the hypothesis is true." In many respects, the most important aspect of this approach is the change it represents in the language used to present forensic science evidence [18].

A general procedure that allows one to estimate LR is given in Fig. 2. A critical first step is careful formulation of the hypothesis that constitutes the "ultimate inference" that is to be tested by the method. Referring to Table 4, tests can generally be classified into one of three categories:

Single hypothesis tests that seek to establish support for a "yes or no" inference. For example, did two samples of agent originate from the same batch of material? Or, was the agent grown on agar plates? One can separate single hypothesis tests into two distinct categories: sample matching and classification.

Multiple hypothesis tests – seek to establish support for a "one of several choices" inference. For example, what growth medium was used to culture the agent?

Calibrations – seek to establish bounds on some parameter associated with a material being tested. For example, was a biological agent produced within a certain time interval in the past?

A well-formed hypothesis is one that can be objectively realized in a set of reference samples that can be subjected to the test. For example, the hypothesis that "the two samples match" would not be well-formed because declaring a match is inherently subjective – i.e. a matter of definition. One can always find differences between two

samples if one looks hard enough, or similarity by increasing the tolerable differences. On the other hand, the hypothesis that "the two samples were drawn from a common batch of material" would be testable, because it is possible to objectively produce test samples that are drawn from same or different batches.

Once a testable hypothesis has been determined, it is necessary to define the *signature,* i.e. the set of molecular, chemical, or physical characteristics that provide the basis for decision (H or $\bar{H}$). In practice, this is often accomplished through an empirical, exploratory study that identifies observable (preferably quantifiable) differences between H and $\bar{H}$ samples. (It is assumed that the measurement process for characterizing the signatures has undergone prior analytical validation, and has been codified as a standard operating procedure (SOP) before the inferential validation study is initiated.) Based on the signature, one then defines an *objective metric for decision,* i.e. a scalar quantity defined in terms of the signature that is used to decide H or $\bar{H}$. The objectivity of the metric is not strictly necessary, but if subjective criteria for decision are used, then the validation procedure strictly applies only to the operator making the subjective decision, not the method in general. Table 5 provides some examples of signatures and possible metrics for various anthrax powder assays that have been discussed in the literature.

In addition to careful hypothesis formulation, careful consideration of the population over which the test applies is essential. The sample set used to perform inferential validation should be "representative" of the population of samples for which the inference is intended, meaning that it is not a biased sampling of members of that population [23]. This follows from the general principle that inferences about the questioned sample based on the properties of a set of reference samples are only valid if all samples were drawn from the same population. Therefore understanding the relevant population and choosing a sampling strategy are key questions that arise in executing the processes outlined in Fig 2.

Two important general observations can be made about the concept of a "population" from which reference samples are drawn:

First, the relevant population may be *real* or *virtual*. For the analysis of materials like fibers or drugs, samples can be drawn from a real population (i.e. materials that already exist) that is generated by commercial manufacturing activities. In contrast, biological agents are clearly not manufactured continuously in quantity, so the "population" of interest is actually determined by the set of *possible* manufacturing processes that could be used to make them. Sampling from this virtual population necessarily involves simulating the diversity in manufacturing methods by using "representative" recipes and laboratories to make reference samples. On the other hand, suppose we wish to validate an antibody assay that is intended to provide evidence that a person received vaccination for anthrax. Clearly the population is real: humans who have and have not received the anthrax vaccine.

Second, whether the population is virtual or real, it is ultimately defined by the types of variation one could expect among real samples. For example, in the case of chemical and physical analysis:

- The exact method of growth and production of an agent
- The exact source of materials used in the production process
- The temperature and humidity conditions under which an agent might have been stored prior to dissemination

or, in the case of the vaccination assay:

- The immune system condition, health, and treatment history of the suspect that a blood sample was drawn from
- The type and formulation of the vaccine that might have been administered.

For a method to be applicable to a questioned sample for which factors like these are not known, the set of samples used for validation must reflect an unbiased selection from a population in which those factors are allowed to vary over their naturally occurring

ranges. Thus, as a prelude to any validation exercise it is necessary to consider the possible factors that could affect the relationship between the measured value of the metric and the hypothesis in question but can not be controlled, and would not be known about a questioned sample.

Once the population is defined, the next critical element of an inferential validation study is to develop a *sampling frame* that adequately represents the population. (A frame is basically a list or tabular representation of actual members of the population that could be sampled [23].) Obviously, the sampling frame should include samples that conform to the hypothesis H and its complement $\overline{H}$, which can be thought of as two sub-populations within the larger population of possible samples. Individual samples are drawn randomly from this list and characterized according to the SOP for the analytic method under study. The metric is computed for each sample, and the end result of the characterization process is two sets of metric values, one from H samples, and one from $\overline{H}$ samples, with their associated probability distributions.

Fig. 3 is a notional representation of distributions of metric values observed for the H and $\overline{H}$ subpopulations, displayed as histograms. A standard way to express the performance of an assay over a population of samples is the receiver-operating characteristic (ROC) curve, which can be constructed in a straightforward way from the metric value distributions [24-26]. Fig. 4 is the ROC curve representation of the data in Fig. 3. Once the population has been characterized this way, the slope of the ROC curve can be used to estimate the likelihood ratio LR using the process illustrated by the dashed arrows in Fig. 4. When a new sample is encountered, it is characterized using the same SOP and the metric value is calculated from the measurement(s). That metric value corresponds to a location on the reference ROC curve. If it lies on the rising part of the curve, the slope (LR) is greater than 1, and the observed metric provides support to the hypothesis H. In this notional example, metric values smaller than 1 favor H, while values larger than 1 favor $\overline{H}$ since $LR < 1$.

The degree of separation between the distributions of metric values for the two sub-

populations is reflected in the steepness of the slope in the ROC curve. If the two sub-populations do not overlap at all, the ROC curve is "perfect", with an infinitely steep slope for values of the metric smaller than the highest value found in the H population. If the two sub-populations are fully overlapping, the resulting ROC curve would have a slope of 1 and the test would have no inferential power regardless of the metric value.

There are several advantages of adopting the ROC/LR framework for inferential validation studies and expert testimony on interpretation. It uses an accepted non-parametric method for interpreting evidence that passes muster with modern evidence scholarship [18]. It avoids the implicit or explicit assumption of prior odds, which may pose problems in some courts [27]. Arguments about the interpretation of assay results based on the ROC/LR framework are likely to center on population, frame, and sampling issues, just as they did for human DNA forensics during its early phases [28]. The issues of population definition and sampling bias are also familiar in a number of other contexts where critical decision-making is dependent on test results, including clinical testing and medical diagnosis [25,26]. Thus, the ROC/LR approach is a generally accepted methodology for scientific inference.

Challenges to population definition generally speak to the weight of the evidence, not admissibility, as long as the bias that might be introduced is not overwhelming, or deliberate. However, the nature of the conceptual source population, and whether the samples used to construct ROC curves are truly representative could clearly be a potential point of contention. It may happen that a study that uses one explicit frame for sampling is called into question when other frames may be reasonably suggested. In this context, inferential validation can be thought of as a multi-phase process as illustrated schematically in Fig. 6.

At an early stage of validation, or under exigent circumstances, only opportunistic or very limited sample sets may be available for testing. The results of such preliminary validation studies may only be useful for generating investigative leads [11]. Test performance is subsequently evaluated on a set of samples drawn from a more carefully

constructed, putatively representative sampling frame and subsequently *validated* on a completely independent set of samples drawn from an independent frame. Standard statistical methods have been developed for testing whether two independent ROC curves or their underlying distributions are drawn from the same underlying population [29]. The results of two or more studies can be combined to make a composite ROC curve that is ostensibly based on a more representative overall population sample. Several cycles of evaluation and validation may occur as our understanding of the structure of the underlying population, and the choice of representative frame evolves. Eventually, reasonable challenges to the population or frame definition must decline, and the ability of the test procedure to provide reliable estimates of the likelihood ratio will become accepted.

The description of the ROC/LR method provided above primarily considered sample matching or classification by single hypothesis tests. The same basic framework also applies to multiple hypothesis tests, although several precautions must be considered. First, the hypotheses encompassed by the test must constitute a complete and non-overlapping set. That is, every possible sample that could be encountered must conform to one member of the set of hypotheses, and only one [30]. Secondly, the inferential power associated with a multiple hypothesis test must be reduced to account for the increased probability of assigning an unknown sample to any particular hypothesis purely by chance [31].

Calibration shares certain technical features with hypothesis testing, although it is a distinct activity. In calibrations we collect a set of data that will allow us to determine the likelihood that a certain parameter of a questioned sample lies within a certain range, based on a measurement (or measurements) of some other property. The use of calibration curves in analytical validation is well known [7,8]. In the context of inferential validation, calibration "curves" (more accurately scatterplots) are constructed by unbiased sampling over the population of sample types, just as ROC curves or other likelihood estimators are for hypothesis tests. The quantitative evidence extracted from a calibration curve can also be expressed as a likelihood ratio [9]. Like hypothesis testing

assays, inferential calibrations are validated by independent re-sampling of the sample population.

4. Application to chemical and physical analysis of biological agents

Tables 1 and 5, and the discussion in section 3 have provided some examples of chemical and physical analysis methods that might be used to infer relationships between two agent samples, or whether certain materials or process steps were involved in their manufacture. This section briefly describes some considerations about framing and sampling the population of manufacturing methods for biological agents for inferential validation studies of chemical and physical methods. Much of this discussion is based on previous experience with validating sample matching tests based on elemental analysis of agents [32,33].

The composition and morphology of a bioagent like *Bacillus anthracis* (Ba) is the end result of the end-to-end process used to produce it. Certain steps influence the overall composition of the agent through the addition or removal of certain substances, and certain steps influence the physical form of the material. The relevant population for evaluating and validating chemical and physical analysis methods is therefore the set of materials that could be generated by any growth and preparation method that might be used to generate a bioagent, using starting materials from any potential sources. Thus, the population of biological agents is an imaginary construct, and the problem is how to generate a set of real samples that adequately provides a statistically representative sample of this imaginary space of possibilities. Moreover, this "population" must be sampled in an unbiased way, capturing all sources of possible variation: Batch-to-batch variation in the same laboratory, laboratory-laboratory variation in executing the same nominal process, and vendor-to-vendor variation in starting material properties.

It is clear from this last requirement that a proper reference sample set for generating ROC curves would involve multiple laboratories making multiple batches of an agent

using multiple processes. How the laboratories and processes are chosen is an important aspect of experimental design, because this choice must be representative of the kinds of laboratories and processes from which case samples are likely to originate. The key is to establish an objective *frame* that represents the population, i.e. a specific list of all the members of the population, and then to use a random selection process to perform the sampling.

One very general frame is illustrated in Fig 6, where each end-to-end process is broken down into unit process steps such as growth, separation of the microbe from the growth medium, washing, drying, milling, and combining with additives. The sampling frame is effectively the list of all combinations of unit processes that plausibly result in an end product. For preparation of a toxin such as ricin, a similar matrix can be constructed with columns defined by the unit process steps appropriate to the particular toxin. For each unit process, there are a number of options, including the "null" option in which that particular unit process is not carried out. Sampling from this frame would involve randomly choosing a variant for each unit process step to create an end-to-end process, then choosing a laboratory at random to execute it. It should be noted that different laboratories might implement a particular unit process in a slightly different way, or use materials from different sources, and this variance can be captured by executing the process at multiple laboratories.

An alternative to the unit process frame is based on the observation that methods for making bacterial preparations are usually communicated as end-to-end recipes. Thus, a valid frame would be a list of all known end-to-end processes that have been used in the past. This is clearly a subset of the possible processes generated by the unit process frame, but arguably captures the most probable processes. Note that both the unit process and end-to-end process frames explicitly connect the validation process with intelligence about terrorist interests and state program practice. Biological agent manufacturing information that a criminal or terrorist might use can come from many sources. This includes material derived from open sources such as recipes provided by underground cookbooks and internet sites, relevant knowledge from the open scientific literature, and

inadvertent leaks of sensitive (but often inaccurate) information that are published in the news media. In some instances intelligence collection efforts may uncover information about the technical knowledge possessed by particular terrorist groups or foreign BW programs. Both of the frames discussed here require periodic updating, and always leave open the question of whether there may be important but unknown sub-populations that have not been sampled.

Assuming that a set of processes and executing laboratories have been randomly chosen from a suitable frame, partial factorial sampling designs can be used to reduce the number of samples to a reasonable value (to control costs.) An example of a design involving 3 processes and 3 laboratories is shown in Fig. 7. The symmetric design helps ensure that the reduction in sample number does not introduce bias. The partitioning of the total number of samples per laboratory among the processes executed by each laboratory represents a degree of freedom that can be optimized for certain tests.

Such designs have been executed for validation exercises involving sample matching and other assays, for the population of bench-top scale processes for producing dry spore agent preparations (using non-pathogenic *B. anthracis* surrogates [33].) Different frames have been constructed and sampled in order to examine the sensitivity of the resulting ROC curves. Preliminary findings indicate that the sampling frames that were discussed above provide a reasonable basis for defining the population, and the method for constructing the sample sets is defensibly unbiased. The resulting sample set provides a useful library for other studies.

5. Concluding remarks:

The ROC/LR method represents a transparent and straightforward approach to inferential validation that uses mainstream statistical concepts and leads naturally to an interpretation of microbial forensic data that does not overstate its probative value. This approach also makes it easy to compare two methods designed for the same purpose or to

combine the results of two independent analyses using orthogonal methods. Although the effort to apply it systematically has only begun recently, it can be applied to a large number of microbial forensic assays. Wider adoption of this methodology will help assure that the interpretation of microbial forensic evidence will meet modern scientific standards.

References

1. National Research Council, Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison, Forensic Analysis: Weighing Bullet Lead Evidence, (National Academies Press, Washington DC, 2004).

2. National Research Council, Committee on Identifying the Needs of the Forensic Science Community: Strengthening Forensic Science in the United States: A Path Forward, (National Academies Press, Washington DC, 2009).

3. Scientific Working Group on Microbial Genetics and Forensics (SWGMGF) "Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work", June 20, 2003. *Science* Supporting Online Material, doi:10.1126/science.1090270.

4. Imwinkelried, E.J., *The Methods of Attacking Scientific Evidence*, 4th Ed. (LexisNexis, 2004).

5. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

6. Federal Rules of Evidence, December 31, 2004.

7. John Mandel, *The Statistical Analysis of Experimental Data*, (Dover Publications, 1984).

8. Principles and Practices of Method Validation, A. Fajgelj and A. Ambrus, eds., (Royal Society of Chemistry, 2000)

9. S. Velsko, "Validation Strategies for Microbial Forensic Analysis of Biological Agents: Beyond Sample Matching", Lawrence Livermore National Laboratory Technical Report UCRL-TR-229944, April, 2007.

10. Budowle B, Johnson MD, Fraser CM, Leighton TJ, Murch RS, and Chakraborty R, "Genetic Analysis and Attribution of Microbial Forensic Evidence", Microbiol. Mol. Biol. Rev. 2006; 70(2):233-254.

11. Schutzer SE, Keim P, Czerwinski K, and Budowle B, "Use of Forensic Methods under Exigent Circumstances Without Full Validation", Science Translational Medicine, 2009; 1(8):1-3.

12. M.J. Saks and J.J. Koehler, "The Coming Paradigm Shift in Forensic Identification Science", Science **309**, pp.892-895, (2005).

13. National Research Council, Committee on DNA Forensic Science, "The Evaluation of Forensic DNA Evidence", (National Academy Press, Washington, D.C. 1996).

14. National Research Council, Committee to Review the Scientific Evidence on the Polygraph, "The Polygraph and Lie Detection", (National Academy Press, Washington,

D.C. 2003).

15. M.M. Houck, "Statistics and trace evidence: The tyranny of numbers", Forensic Science Communications **1**(3), October 1999.

16. R. Lempert, "Modeling Relevence", Michigan Law Review Vol. 75, pp.1021-1057, (1977)

17. D.H. Kaye and J. J. Koehler, "The Misquantification of Probative Value", Law and Human Behavior 27, 645-659, (2003).

18. B. Robertson and G.A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, (J. Wiley and Sons, Ltd, 1995.)

19. Wunschel DS, Colburn HA, Fox A, Fox KF, Harley WM, Wahl JH, and Wahl KL, "Detection of agar, by analysis of sugar markers, associated with *Bacillus anthracis* spores, after culture", J. Microbiol. Methods 2008; 74:57-63.

20. Brewer LN, Ohlhausen JA, Kotula PG, and Michael JR, "Forensic Analsis of bioagents by X-ray and TOF-SIMS hyperspectral imaging", Forensic Science International 2008; 179:98-106.

21. Whiteaker, JR, Fenselau, CC, Fetterolf, D, Steele, D and Wilson, D, "Quantitative Determination of Heme for Forensic Characterization of Bacillus Spores Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry", Anal. Chem. **76**, 2836-2841, (2004).

22. Kreuzer-Martin HW, and Jarman KH, "Stable Isotope Ratios and Forensic Analysis of Microorganisms", Appl. Environ. Microbiol. 2007; 73:3896-3908.

23. William E. Deming, Some Theory of Sampling (Dover Publications, Inc., New York, 1966)

24. Krzanowski WJ and Hand DJ, ROC Curves for Continuous Data, (CRC Press, Boca Raton, 2009).

25. M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford Statistical Science Series 28; Oxford University Press, New York, 2003)

26. NCCLS. "Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristics (ROC) Plots; Approved Guideline. NCCLS Document GP10-A, December, 1995.

27. Good PI, Applying Statistics in the Courtroom, (Chapman and Hall/CRC, CRC Press LLC, Boca Raton, FL, 2001).

28. National Research Council, Committee on DNA Forensic Science, "The Evaluation of Forensic DNA Evidence", (National Academy Press, Washington, D.C. 1996)

29. Kester, A.D.M, and Buntinx, F., "Meta-Analysis of ROC curves", Medical Decision Making, **20**, pp.430-439, (2000).

30. *Classification, Estimation, and Pattern Recognition*, by T.Y. Young and T.W. Calvert (American Elsevier Publishing Company, Inc., New York, 1974.

31. Shafer JP, "Multiple Hypothesis Testing", Ann. Rev. Psych. 1995; 46:561-584.

32. Velsko, S., "Bioagent Sample Matching using Elemental Composition Data: An Approach to Validation", Lawrence Livermore National Laboratory Report UCRL-TR-220803, April, 2006.

33. Velsko SP, Weber P, Ramon CE, Lindvall RE, Davisson ML, and Robel M, "Bioagent sample matching using elemental composition data", Lawrence Livermore National Laboratory Report LLNL-TR-419683 September 30, 2009.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Tables for *Inferential validation and evidence interpretation* (Velsko)

Table 1. Assays and inferences in microbial forensics

Test or assay results	*Ultimate inference*
Elemental profiles of two agent samples	**The two agent samples were made by the same (different) method(s)**
Carbon 14 content of an agent sample	**The agent was produced later than a certain date**
Presence of certain organic compounds	**The agent was made using certain materials or methods**
Genetic sequences of two bacterial isolates A and B.	**Isolate A could have been derived by culturing isolate B.**

Table 2. Relationship between analytical and inferential validation.

Daubert criterion	*Type of expert witness*	*Type of validation*	*Statistical metrics*	*Assurance value*
Reliability	Reporting	Analytical	Precision Accuracy Reproducibility	The technique produces consistent, objective results with known precision and traceable accuracy
Relevance	Interpreting	Inferential	ROC curves Likelihood ratios	The result supports the expert's inference

Table 3. Required types of validation for the four validation categories.

	Analytical	Inferential
Preliminary	Yes	Yes
Developmental	Yes	Yes
Internal	Yes	No
External	Yes	No

Table 4. Most assays can be placed in one of 4 categories

	Type of test	Purpose of test	Examples
	Sample matching	To establish that two agent samples originate from the same batch of material, or were made by the same process	Did the Leahy and Daschle samples come from the same batch?
	Classification: Single hypothesis	To establish that a certain material or that a certain process condition was used in the manufacture of the agent	Was the sample grown on agar plates?
	Classification: Multiple hypotheses		Which growth medium was used?
	Calibration	To establish bounds on some parameter associated with the agent	How old is the sample?

Table 5. Examples of signatures and metrics for some notional anthrax powder assays based on published work.

Assay/Test	Signature	Metric
Assay for presence of residual agar [19]	Mass spectral peaks at relevant m/z values	Ion counts at each m/z value
Assay for presence of added silica [20]	X-Ray emission (EDX) spectrum for Si and O	Peak areas
Assay for presence of residual heme [21]	MALDI mass spectral peaks	Sum of peak heights
Sample matching using isotopes [22]	Stable isotope ratios for C,N,O and H	Euclidian distance between isotopic δ values for two samples

Figures

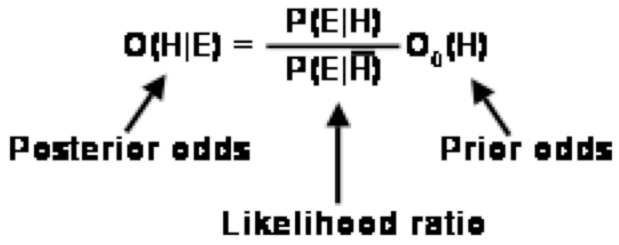

Figure 1. Basic equation for interpreting forensic evidence.

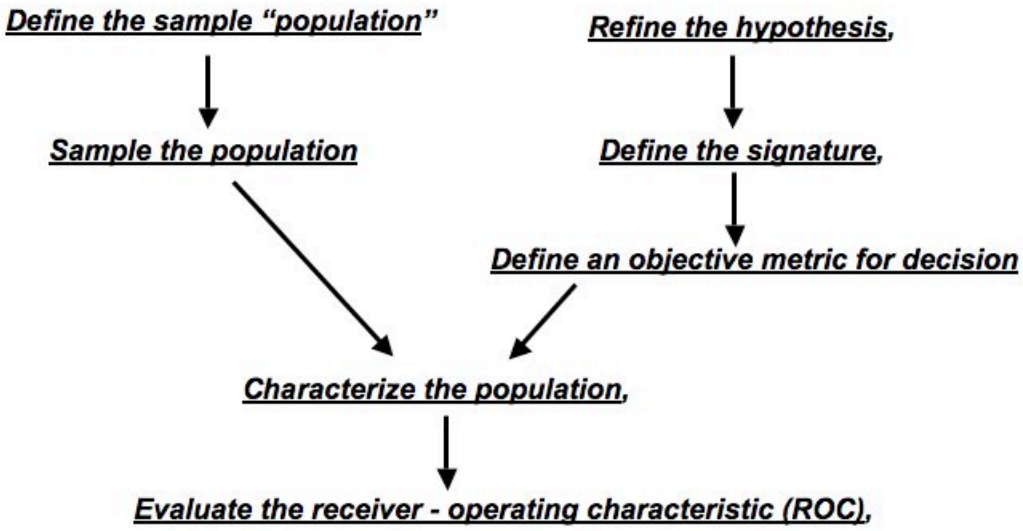

Figure 2. Steps for a generic inferential validation study.

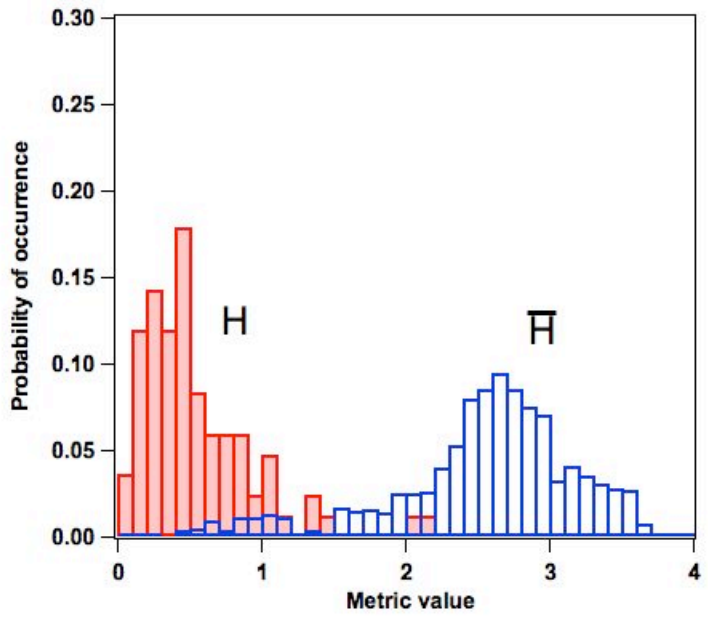

Figure 3. A notional example of a histogram of metric values resulting from characterizing H and $\bar{H}$ subpopulations.

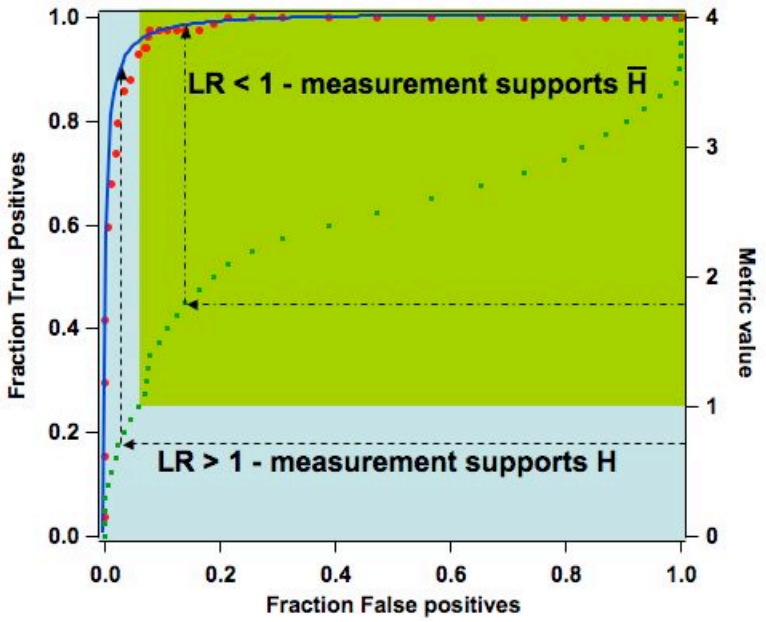

Figure 4. The ROC curve corresponding to the data in Fig. 3. Red dots: empirical ROC curve; Blue curve: fitted ROC curve; Green dots: metric values; The green and blue regions demarcate zones of positive and negative support for the hypothesis H.

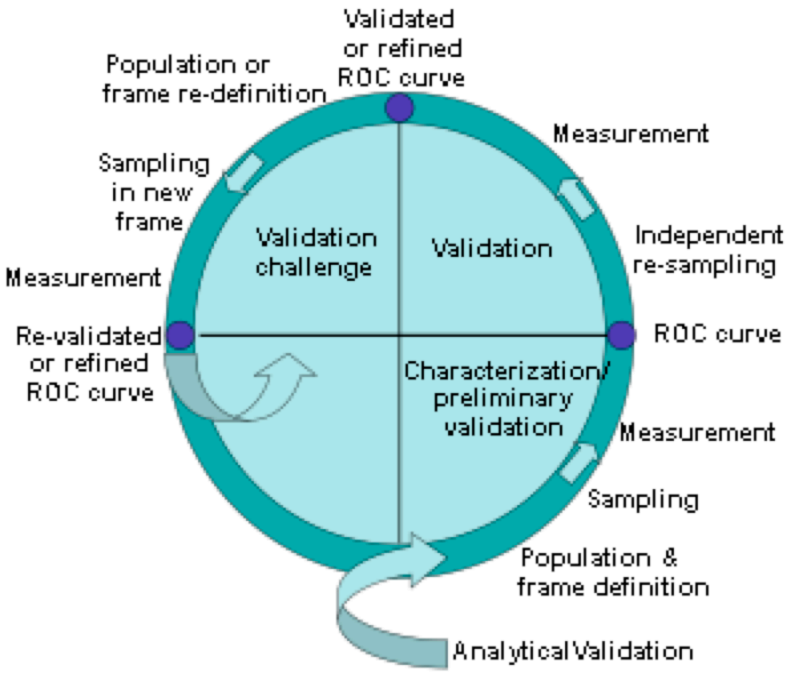

Figure 5. The ROC/LR approach defines a cycle for continuous improvement

Growth	Separation	Washing	Drying	Milling	Additives
	∅	∅	∅	∅	∅
G_1	S_1	W_1	D_1	M_1	A_1
G_2	S_2	W_2	D_2	M_2	A_2
G_3	S_3	W_3	D_3	M_3	A_3
⋮	⋮	⋮	⋮	⋮	⋮

Figure 6. Unit process decomposition of biological agent production. An end-to-end process draws a process variant from each column.

	Lab 1	Lab 2	Lab 3	# batches per process
Process 1	Batch 1 Batch 2 Batch 3	Batch 5 Batch 6	Batch4	6
Process 2	Batch 4	Batch 1 Batch 2 Batch 3	Batch 5 Batch 6	6
Process 3	Batch 5 Batch 6	Batch 4	Batch 1 Batch 2 Batch 3	6
# batches per lab	6	6	6	Total # of batches = 18

Figure 7. A 3 x 3 partial factorial design for sample production.